AptaCluster - A Method to Cluster HT-SELEX Aptamer Pools and Lessons from its Application

Jan Hoinka[1,⋆], Alexey Berezhnoy[2,⋆], Zuben E. Sauna[3], Eli Gilboa[2,⋆⋆], and Teresa M. Przytycka[1,⋆⋆]

[1] National Center of Biotechnology Information, National Library of Medicine, NIH, Bethesda MD 20894, USA przytyck@ncbi.nlm.nih.gov

[2] Department of Microbiology & Immunology, University of Miami Miller School of Medicine, Miami, Florida 33101, USA

[3] Laboratory of Hemostasis, Division of Hematology, Center for Biologics Evaluation and Research, Food and Drug Administration, Bethesda, Maryland, USA

Abstract. Systematic Evolution of Ligands by EXponential Enrichment (SELEX) is a well established experimental procedure to identify aptamers - synthetic single-stranded (ribo)nucleic molecules that bind to a given molecular target. Recently, new sequencing technologies have revolutionized the SELEX protocol by allowing for deep sequencing of the selection pools after each cycle. The emergence of High Throughput SELEX (HT-SELEX) has opened the field to new computational opportunities and challenges that are yet to be addressed. To aid the analysis of the results of HT-SELEX and to advance the understanding of the selection process itself, we developed AptaCluster. This algorithm allows for an efficient clustering of whole HT-SELEX aptamer pools; a task that could not be accomplished with traditional clustering algorithms due to the enormous size of such datasets. We performed HT-SELEX with Interleukin 10 receptor alpha chain (IL-10RA) as the target molecule and used AptaCluster to analyze the resulting sequences. AptaCluster allowed for the first survey of the relationships between sequences in different selection rounds and revealed previously not appreciated properties of the SELEX protocol. As the first tool of this kind, AptaCluster enables novel ways to analyze and to optimize the HT-SELEX procedure. Our AptaCluster algorithm is available as a very fast multiprocessor implementation upon request.

1 Introduction

Aptamers are short, (∼20 to ∼100 nucleotides) synthetic, single-stranded (ribo)-nucleic molecules that can be generated to bind specifically to molecular targets. These binding targets can vary from small organic molecules [1], through proteins and protein complexes [2], to viruses [3], and cells [4]. Aptamers have high structural stability over a wide range of pH and temperatures making them ideal

⋆ First Authors

⋆⋆ Corresponding Authors

reagents for a broad spectrum of in-vitro, ex-vivo, and in-vivo applications [5]. A pegylated aptamer that inhibits binding of Vascular Endothelial Growth Factor (VEGF) to the VEGF receptor (Macugen ®) is approved for the treatment of age-related macular degeneration [6]. Aptamers can also be used to monitor small changes in the conformation of proteins, a property that can be utilized for detecting the effect changes in the manufacturing process or during the development of generic versions of protein-therapeutics [7].

Aptamers are experimentally identified through a procedure known as Systematic Evolution of Ligands by EXponential Enrichment (SELEX) [8]. The traditional SELEX procedure iterates over five basic steps which together define one selection cycle: incubation, binding, partitioning and washing, target-bound elution, and amplification (Fig. 1). The process starts with a single-stranded (ribo)nucleic acid sequence library of, typically, 10^{15} random sequences of fixed length flanked by constant primer sites to aid amplification. Each random sequence permits the molecule to fold into a unique 3D shape or conformation. At the start of each cycle, such a RNA/ssDNA pool is incubated with a target of interest. Due to the large number of unique sequences in the library, the probability of at least some aptamer molecules to bind the target with specificity and affinity is quite high. At the end of each cycle, low affinity binders are removed from the solution whereas bound aptamer molecules are eluted and amplified, forming the input for the next round. Eventually, only molecules that bind the target with high affinity remain. The aptamer molecules thus selected for high affinity and specificity are then individually evaluated experimentally and optimized for specific properties, such as size or stability, depending on the intended application. The experimental optimization is often assisted by computational analysis. Such analysis includes finding minimum free energy secondary structures and the identification of sequence motifs common to the final pool of aptamers. Recently, Hoinka et al. developed AptaMotif, a computational method for the identification of sequence-structure motifs in SELEX-derived aptamers [9].

New sequencing technologies have revolutionized the SELEX protocol by allowing deep/next-generation sequencing of entire aptamer pools ([10], Fig. 1). This extension, the so-called HT-SELEX, holds the promise for greatly accelerating aptamer discoveries and expanding their applications. For example, in the special case where the target molecule is a transcription factor, a variant of HT-SELEX designed for double-stranded DNA aptamers has been successfully used to uncover transcription factor binding motifs [11–13].

Traditionally, the SELEX process has been treated as a black box and only a handful of binders elucidated in the last cycle were sequenced. In contrast, sequencing of earlier pools using HT-SELEX provides the opportunity to uncover potential binders that might otherwise have been lost in later steps of the selection process. More importantly, by analyzing the relative changes of consecutive

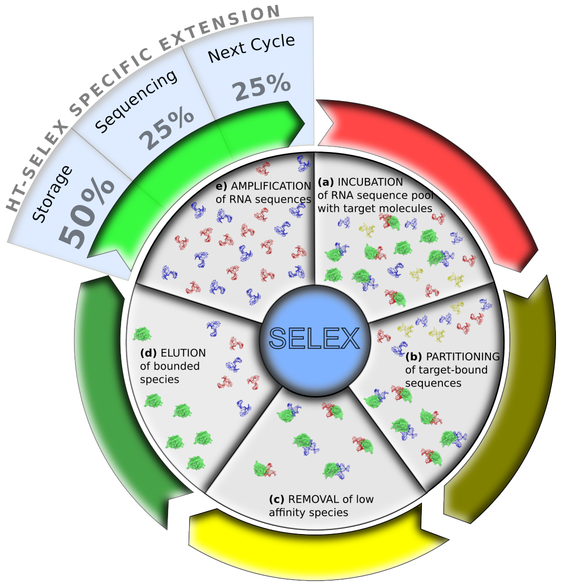
HT-SELEX SPECIFIC EXTENSION
Storage
50%
Sequencing
25%
Next Cycle
25%
e) AMPLIFICATION of RNA sequences
(a) INCUBATION of RNA sequence pool with target molecules
SELEX
(b) PARTITIONING of target-bound sequences
(c) REMOVAL of low affinity species
(d) ELUTION of bounded species

Fig. 1. The SELEX procedure iterates over five basic steps incubation, binding, partitioning and washing, target-bound elution, and amplification. Traditionally, only the binders elucidated in the last cycle were sampled and examined. The HT-SELEX includes sequencing of the final and intermediate selection pools.

selection rounds of properties such as sequence diversity and mutation rates, the method provides an unprecedented opportunity to gain deeper insights into the selection process *per se*. Thus, HT-SELEX coupled with computational assessment of the relation between sequences has the potential to trace the dynamics of the selection process and the rational selection of aptamers with desired properties making the SELEX process more rapid as well as more efficient.

Despite the success of HT-SELEX for drug design, efficient computational tools that exploit and encompass data from all sequenced rounds, therefore elucidating the selection process from the initial pool to the final cycles, have yet to be developed. Computational processing of HT-SELEX data is currently largely based on simple counting of aptamer species in the final round of selection, frequently discarding low-frequency species from the analysis, and choosing the sequences that occur in high counts for further investigation [10]. In addition, a small number of most frequent sequences from the final selection round might be used as seeds for similarity searches. The underlying postulate of these methods is that the best predictor of binding affinity is the frequency at which a particular aptamer occurs in a pool. While these approaches might be suitable for candidate identification, they lack the ability of providing insight into the mechanisms governing the selection process itself. Note that as the selection progresses, low affinity binders (Fig. 2 high z-coordinate) are eliminated from the pool leaving

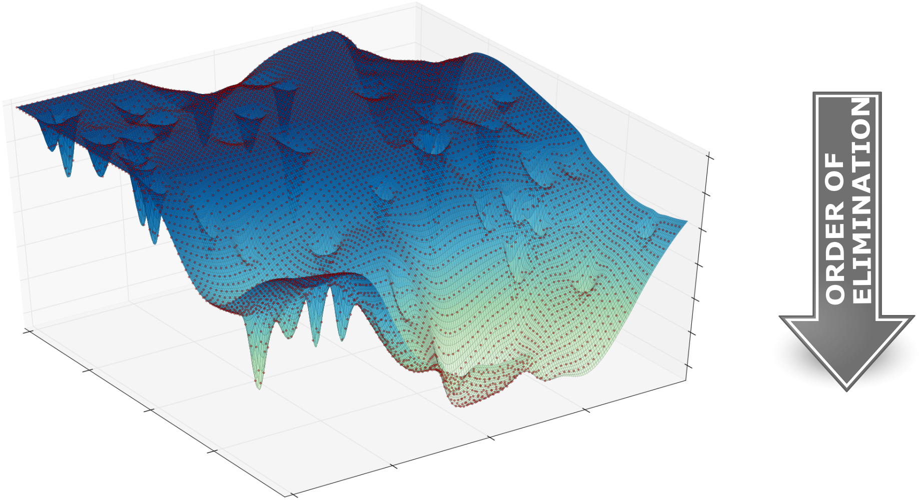

Fig. 2. A visualization of the aptamer landscape probed by the SELEX protocol. The surface represents all possible aptamers of fixed length and the red dots represent aptamers used in the initial pool. The distance on the surface is a conceptual projection of sequence similarity. Multiple local minima correspond to groups of aptamers that bind to the different areas of the targets surface or to the same region but are related by structure rather than sequence similarity.

aptamers that sample local minima of binding energy. It is therefore expected that clustering of aptamers in consecutive cycles should provide valuable information about the selection process and should allow for the delineation of the entire aptamer landscape probed by the SELEX protocol. Hence, our primary objective is to cluster aptamers in all rounds of selection according to their sequence similarity. This task however could not be accomplished with previous clustering algorithms due to the enormous size (2-50 Million sequences per cycle) of the data set generated by high throughput sequencing, especially for early rounds of selection which feature a high degree of unique sequences ($\geq 90\%$). To address this challenge, we developed a novel approach, AptaCluster, capable of efficiently clustering entire aptamer pools.

Several sequence similarity measures are commonly found in clustering methods, of which the Hamming and Levenshtein (edit) distances are most prominent. However, full-scale clustering approaches are computationally untrackable for HT-SELEX data. Therefore we use the randomized dimensionality reduction technique, known as locality-sensitive hashing (LSH) [14], to implicitly approximate an upper bound to the edit distance for each sequence pair without the need of exhaustive pairwise comparison. In the subsequent step, we eventually compute precise sequence distances based on k-mer counting between pairs of aptamers below this bound, while the remaining distances are not relevant and might be arbitrarily assumed to be infinity.

We applied AptaCluster to analyze the results of the HT-SELEX experiment that we preformed using Interleukin 10 receptor alpha chain (IL-10RA) as the target molecule. IL-10 is considered to be a master regulator of immunity to infection and is an important therapeutic molecular target [15]. We preformed 5 cycles of HT-SELEX with a 40nt variable region, sequencing the samples of pools 2-5. AptaCluster has enabled us to analyze the results of HT-SELEX, revealed interesting properties of the selection landscape, and allowed for a better understanding of the HT-SELEX experiment. AptaCluster scales very well with data size. While the sequenced pools in our IL-10RA HT-SELEX experiment varies between 2 and 4.5 Million aptamers, we have applied AptaCluster to much larger pools of more than 20 Million sequences in the context of whole-cell HT-SELEX (data not shown) without loss of noticeable performance.

2 The AptaCluster Algorithm

Our approach is centered around a randomized dimensionality reduction technique, known as locality-sensitive hashing (LSH) [14]. First, a compressed representation of the data set is constructed by reducing the pool to non-redundant species and their corresponding frequency counts. We then apply a user-defined number of randomized locality-sensitive hash functions to the data set in order to distinguish sequence pairs that are potentially similar from those that are, with very high probability, not similar. Each function operates by selecting a small number of nucleotide positions from each aptamer and treats the substring, resulting from the concatenation of these bases, as input for the hashing procedure. Hence, aptamers with highly similar primary structure are likely to fall into the same group whereas dissimilar sequences rarely produce identical hash values. In the third step, the actual clustering step, we compute precise sequence distances between aptamers of identical hash value, while the distances between the aptamers never encountered in the same group are set to infinity. To accelerate the clustering, AptaCluster relies on a similarity measure based on k-mer counting. Thus the algorithm preforms three main steps outlined below. Relevant implementation details and the parameters used throughout this study can be found in the Methods section.

Dataset Compression Data compression is achieved by using a hash map in which the keys correspond to the species in the pool and the values correspond to their respective frequency counts which can be done in $\mathcal{O}(N)$ time. In the following, let $s = (s_i)_{i=1}^{l}$ be an aptamer of of length l defined by the sequence of nucleotides s_i over the alphabet $\Omega = \{A, C, G, T\}$ where the index i corresponds to the i^{th} position of the aptamer. Furthermore, we define $S = \{s^j \in P \parallel s^j \neq s^k \; \forall j, k \in [1, \ldots, |S|] \wedge \sum_{j=1}^{|S|} m(s^j) = N\}$, where $m(s^j)$ corresponds to the frequency of s^i, as the keys of the hash map, i.e. the set of unique aptamers for pool P.

Filtering using Locality Sensitive Hashing LSH is based on the idea that data points that are close in high dimension, after applying a probabilistic dimensionality reduction and using the reduced representation as the input to a hash function, are likely to obtain the same hash value and hence fall into the same bucket [16].

AptaCluster exploits this property by treating each sequence $s^j \in S$ as an l-dimensional vector and reducing this vector into d dimensions ($d < l$). This is done by generating a set I_d of d randomly sampled indices $i \in [1, \ldots, l]$ and, for each sequence s^j, only selecting those nucleotides s_i for which $i \in I_d$ as input for the hashing procedure. Hence, the more similar the primary structure of a set of aptamers, the higher the probability that they will produce the same mapping. Similarly, the choice of d controls the minimal degree of similarity between the members of each partition since these are guaranteed to differ in at most $l - d$ positions. In other words, our approach implicitly computes an upper bound to the edit distance. We iteratively improve this upper bound by repeating this procedure a user defined number of times, each time using a different hash function. With sufficient number of of iterations, if two sequences never fall into the same bucket they are assumed to be dissimilar with very high probability. The iterative computation of the upper bound is performed as follows. Let $d^k_{lsh}(s^1, s^2)$ be the upper bound computed after the k^{th} iteration and let $L^k(s)$ be the value of the k^{th} hash function for sequence s. We assume that, by default, we have for all pairs $d^0_{lsh}(s^1, s^2) = \infty$. Then

$$d^k_{lsh}(s^1, s^2) = \begin{cases} l - d & L^k(s^1) = L^k(s^2) \\ d^{k-1}_{lsh}(s^1, s^2) & L^k(s^1) \neq L^k(s^2) \end{cases} \tag{1}$$

Clearly, only the assignment in the first line needs to be executed. To define $L^k(s)$, for each iteration k we randomly select a mapping h from a family of functions

$$F = \{h : \mathbb{N}^l \to \mathbb{N}^d \parallel h(I) = I_d\} \tag{2}$$

where $I = (1, \ldots, l)$ represents the nucleotide positions of an aptamer of size l, and apply the function

$$L = \{\Omega^l \to \Omega^d \parallel L(s) = (s_i) \; \forall \; i \in I_d\} \tag{3}$$

to each aptamer s, creating a sub-string $\hat{s}$ comprised of the concatenation of the nucleotides at the positions defined in I_d. Finally, traditional hashing is performed on the set $\hat{S} = \{\hat{s}^i\}, i = 1, \ldots, |S|$. $I_d = (i_0, \ldots, i_d)$ can be efficiently computed as follows: Let $i_0 \in [1, l]$ be a randomly selected index of I and define $x \in [2, l - 1]$ as a random number co-prime to l. Then, the remaining positions can be generated with

$$i_j = (i_{j-1} + x) \mod l, \quad j = 1, \ldots, d - 1 \tag{4}$$

and

$$I_d = (i_j)_{j=0}^{d-1}\,, \quad i_j < i_{j+1}\;\forall\; j \tag{5}$$

corresponds to the sequence of indices after sorting these in ascending order. Using this scheme guarantees that each index in I is selected exactly once and avoids scenarios in which only adjacent positions of the sequence are chosen.

Cluster Extraction Based on the assumption that high-frequency of a sequence in a selection pool is related to its selective advantage due to its binding affinity, we build the clusters iteratively around these high frequency aptamers. We repeatedly choose the highest frequency sequence s not assigned to any cluster, making it a seed of the new cluster. We then we employ a k-mer based distance function [17] to compute the distance of the selected seeds to all other sequences for which the upper bound estimated with LSH was finite and include it in the cluster if d_{kmer} is smaller than a user defined cutoff. In particular,

$$d_{kmer}(s^x, s^y) = \sum_{i=1}^{4k} \left| \frac{X_i}{|s^x| - k + 1} - \frac{Y_i}{|s^y| - k + 1} \right|^2 \tag{6}$$

where X_i and Y_i denotes the number of times the i-th k-mer occurs in sequence s^x and s^y respectively and $|s^i|$ corresponds to the length of the aptamer. Since we compare only sequences that are in the same bucket in at least one iteration, this approach allows us to extract clusters in $\mathcal{O}(N * m * k)$ where m denotes to the maximum number of seed sequences in a bucket which is bounded by the size of the largest bucket generated during LSH.

3 Results of application to HT-SELEX experiment for IL-10RA

We performed 5 rounds of HT-SELEX experiment with Interleukin 10 receptor alpha chain (IL-10RA) as the target molecule. Here, we summarize the insights obtained using AptaCluster.

3.1 Validating Clustering Results

The main advantage of AptaCluster is that to cluster an aptamer pool it does not need to compute the distances between all pairs of sequences but instead uses locality-sensitive hashing to filter out pairs that do not need to be compared. However, the filtering step is heuristic and its outcome might depend on the number of LHS iterations and properties of the dataset. Therefore we started by confirming that the filtering step produces correct results, i.e. that sequences filtered out as not potentially similar are indeed remote from the seed sequences in terms of exact distance. Since the dataset size prohibits an exhaustive computation of all distances, we used 400 aptamers (the 20 most frequent species from the top 20 clusters) and computed their edit distances to all other aptamers.

We then computed the distribution of the distances to the members of the same cluster to the distances to the rest of the aptamers. The former group sampled the sequences whose distances to the reference sequences has been computed and found to be below the clustering threshold. The latter group sampled two types of sequences: the sequences whose distance to the reference sequence has been computed but found to be above the threshold and the sequences filtered out without computing the distance based on our locality sensitive hashing function. The results for all selection cycles are summarized in Fig. 3 for a set of default and relaxed parameters (see Parameters section). The results demonstrate that no sequence that was filtered out using locality sensitive hashing is close to the seed sequences of the clusters. In addition, it also demonstrates that SELEX derived aptamer clusters are well separated. Indeed, relaxing the locality-sensitive hashing based filtering and increasing clustering threshold did not change the clustering results appreciatively (Fig. 3 (b)).

3.2 Distribution of Aptamers within Clusters

Next, we examined the distribution of aptamers within the clusters. Interestingly, we found that the distribution of these frequencies was very skewed (Fig. 4). Except for a handful of highly abundant aptamers, most of the species in a cluster had low frequencies. Such extreme differences in frequencies is consistent with a situation in which most of the cluster diversity can be attributed to mutations caused by Polymerase errors. To test this hypothesis, we investigated whether aptamers with a maximal count of 5 from the top 20 clusters in cycle 5 were also present in the sequenced portion of the selection pool from cycle 2. Indeed, the vast majority of these sequences (99% of singletons, 97% for frequency 5) where absent in this pool (Supplementary Table 1). Note that the sequences introduced by Polymerase errors can be subsequently selected and amplified providing an important source of cluster's diversity. However, due to the late introduction, their frequency count might not correctly reflect their binding affinity.

3.3 Frequency Counts Versus Binding Affinity

It is often assumed that an aptamer sequence's frequency in the pool later cycles provides a good predictor of its binding affinity. Indeed this would be a reasonable expectation under the assumption that the selection process is free of any artifacts, all aptamers are present in the initial pool with the same frequency, and there was no stochastic variability during the above mentioned partitioning. However the realization that a large fraction of sequences in the final pool might have been absent from the initial pool but introduced in a later stage made us to reexamine this assumption. We measured disassociation constant K_d for 30 Aptamers including the most frequent ones. We found that cycle-to-cycle enrichment of aptamer frequencies, i.e. their relative increase in multiplicity, from cycle 4 to 5 is a better predictor of binding than the frequency in the final pool (data not shown). Specifically, taking 125 K_d as a reasonable threshold between

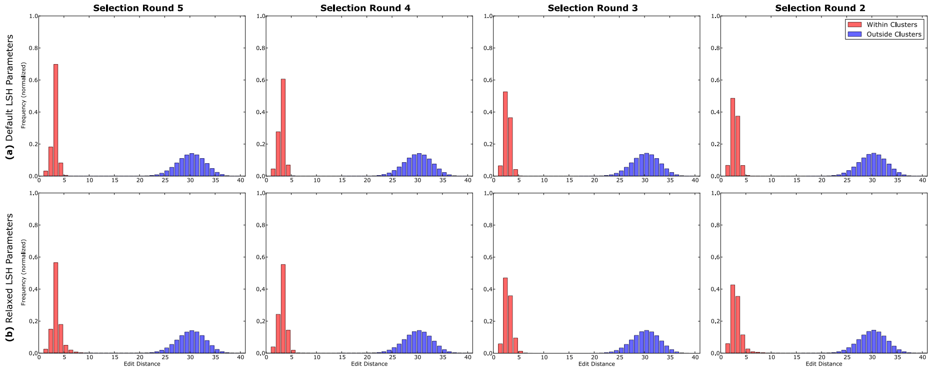

Fig. 3. Distribution of the edit distances between aptamers belonging to a cluster (red) and distances between cluster members and all non-cluster sequences (blue) for selection rounds 2 to 5. Within each of the top 20 clusters, the 20 most frequent aptamers where compared against all other cluster members as well as the remaining aptamers of the pool. (a) Distributions using the defaults parameters of AptaCluster as described in the Parameter section is shown in the top panel. (b) Relaxed parameters as depicted in the bottom panel in which only 40% of the randomized region was sampled during LSH.

binders and non-binders, sorting by cycle-to-cycle enrichment separates binders form non-binders while sorting by frequency leaves these two groups randomly mixed.

In addition to the emergence of new sequences, another source of dissonance between aptamer frequency and its binding potential could also be the differences in their frequencies in the initial pools due to the stochastic nature of partitioning the pool into groups to be used for storage/sequencing/next cycle. Looking at cycle-to-cycle sequence enrichment instead of counts permits a resolution of this problem. However, other artifacts exist that can affect aptamer frequencies as well. In particular, we also tested the K_d values for non IL-10RA specific binding using binding to IgG as proxy for such non-specificity (data not shown). We found for example that cluster with ID 3 has high frequency in cycle 5 but it is not IL-10RA specific.

4 Conclusions and Discussion

Given the great promise of the HT-SELEX approach and rapidly diminishing costs of next generation sequencing, the usage of this method is likely to increase rapidly. Therefore it is imperative that researchers are able to analyze and correctly interpret HT-SELEX results. We have developed a new approach, AptaCluster, that allows for clustering based on primary structure of pools of aptamers sequenced using Hi-Seq technology.

Until now, a typical HT-SELEX analysis was reduced to counting the frequency of each aptamer and using such counts as a predictor of binding affinity. However our results indicate that such counting is actually not as good of a predictor as it has been anticipated. Instead, a predictor that utilizes the dynamics of the cycle-to-cycle enrichment holds greater promise.

Our results of applying AptaCluster to the outcome of the IL-10RA HT-SELEX experiment revealed important properties of the resulting clusters. We found the clusters to be well separated, and typically dominated by one or a few individuals. Relaxing the parameters to allow for larger intra-cluster distances did not change the results significantly. Consequently, sequence profiles of individual clusters were dominated by one or a few of the most abundant sequences. We have also implemented a procedure that enables the tracing of the clusters over consecutive selection cycles and, consistently with the observation above, we found that the clusters' sequence profile did not change much during consecutive selection steps.

The distribution of frequency counts within clusters suggests that cluster diversity is, in a large part, a result of Polymerase errors. The emergence of such Polymerase mutants creates an interesting opportunity to sample around local minima. This is strengthened by the observation that the number of mutations correlates with the frequency of the cluster seeds: the more frequent the seed, the more frequent the mutants. How to design the dynamics of the selection process to optimally utilize these emerging mutants is an open question. One possibility is to replace the typical selection procedure where selection pressure increases in each cycle by an approach that alternates between stronger and weaker selection.

AptaCluster provides a valuable tool which will help us and others to analyze and to optimize the HT-SELEX procedure. It has enabled us to analyze the results of HT-SELEX for IL-10 and allowed for a better understanding of the HT-SELEX experiment. We expect that the properties of the clusters obtained with AptaCluster will vary depending on the experimental details of HT-SELEX protocol in use, the length of the variable region, error rate of Polymerase, and properties of the target. Independently of this expected variability, AptaCluster can be used as the first step towards understanding the aptamer binding landscape, and for the identification of a broad spectrum of potential binders. We point out that AptaCluster is not intended to elucidate complex, indel-containing motifs but rather to operate on sequences of equal length. It is designed to serve as a preprocessing step for approaches to uncover sequence-structure motifs such as the planned extension of our AptaMotif algorithm to high throughput sequencing data [9].

5 Materials and Methods

5.1 Dataset Description

We applied 4 rounds of selection and cDNA generated from round 5 bound fractions as well as RNA recovered from bound fractions at rounds 2, 3 and 4 was amplified and sequenced using Illuminas HiSeq 2500 device with 100-cycle paired-end sequencing protocol (see HT-SELEX Experiment section for the experimental protocol). Aptamers were extracted by aligning the transcribed, inverted sequence of the reverse run to the corresponding forward lane and only retaining those sequences with less then 5 mismatches between the actual primers/tags and the identified primer region. Furthermore we restricted the number of allowed mismatches between the sequences of the forward and reverse lane in the randomized region to four. Mismatches in the randomized region were corrected by choosing the nucleotide with higher Illumina quality score. For the entire experiment, a total of 12895554 sequences where retrieved of which 4621438 species belonged to round 5, 1923823 to round 4, 2181720 to round 3, and 4168573 to round 2. Out of these respectively 617220, 1021668, 1902904, and 3857210 were unique.

5.2 Implementation Details

AptaCluster is currently available as a multi-threaded implementation in C++ using the OpenMP and Boost libraries for its parallel programming operations and hashing procedures, respectively [18, 19]. It features a complete, highly modular pipeline from data input and parsing, over cluster extraction, to result visualization and database storage. We implemented threaded parsers for a number of file formats, including FASTA, FASTQ, and RAW sequence files, both for paired-end and single-end sequencing data as well as automatic multiplexing procedures for separating the individual SELEX rounds when sequenced together. Depending on the number of available CPUs, clustering and distance calculations are performed in parallel for each pool. Cluster families and their evolution from cycle to cycle are currently visualized in HTML format. Finally, the algorithms behavior can be controlled using a configuration file allowing for the assignment of most parameters used for parsing and clustering, among others. We have empirically determined a set of default values, of which the most relevant are discussed below.

5.3 Parameters

For the experiments described in this paper, we performed a total of $r = 10$ iterations of LSH sampling 60% of the randomized region (i.e. $l = 24$). The parameter $d = 4$ is set in terms of the maximal number of point mutations any pair of sequences should have and is converted into the k-mer distance cutoff by sampling a user defined number of aptamers from the pool (10000 by default), artificially mutating that sequence up to d times, and averaging over all d_{kmer}

between these mutants and the wild-type. Furthermore we set $k = 3$ for the computation of d_{kmer} which has shown to give reasonable results for aptamer-sized sequences.

5.4 HT-SELEX Experiment

Selection Details. A DNA template for the selection library was ordered from IDT (Coralville, IA). 1 nM of each N_{40} template (5-TCTCGATCTCAGCGAGTCGTCG -N_{40}-CCCATCCCTCTTCCTCTCTCCC-3) and 5 primer (5-GGGGGAATTCTAATACGACTC ACTATAGGGAGAGAGGAAGAGGGATGGG-3) were annealed together, extended with Taq polymerase (Life Science), and transcribed in vitro using Durascribe (in-vitro transcription) IVT kit (Illumina). The random R0 RNA was purified by denaturing PAGE and, after preclearing with human IgG-coated (Sigma) beads (GE Healthcare), used for in-vitro selection. 1 nM of R0 RNA was used in a first round of selection to coincubate with 0.3 nM of bead-bound human IL-10RA-Fc fusion protein (Novus Biologicals) in 100 mM NaCl selection buffer. After washes, a recovered bound RNA fraction was reverse transcribed using the cloned AMV RT kit (Life Science). cDNA was amplified by either emulsion or open PCR using Platinum Taq PCR kit (Life Science) as described below. The DNA template was used to IVT RNA for the next round. During subsequent rounds, amount of protein was reduced 25% each time, while concentration of NaCl was gradually increased to 150 mM.

Emulsion PCR. cDNA was amplified using Platinum Taq PCR kit with addition of 10% PCRx enhancer solution and following primers: 5-GGGGGAATTCTAAT ACGACTCACTATAGGGAGAGAGGAAGAGGGATGGG-3 and 5-TCTCGATCTCAGCGAGTCGTCG-3. After preparing the master mix PCR reaction solution, it was separated to 100 μL aliquots and each aliquot was mixed with 600 μL ice-cold oil fraction assembled from components supplied with emulsion PCR kit (EURx) according to manufacturers instructions. Water and oil mixture was emulsified by 5 vortexing at +4C and amplified in standard PCR machine for 25 cycles. Control open PCR reaction was carried with aqueous phase only for 16 cycles.

Preparing Libraries for HTS. After 4 rounds of selection, 3 nM of RNA was prepared for round 5. The RNA was pre-cleared using IgG-coated beads and separated into three identical aliquots. Each aliquot was incubated with either human IL10RA protein, murine IL10RA protein or human IgG. After standard washes, bound RNA fraction was extracted from beads and reverse transcribed as described previously. A cDNA generated from round 5 bound fractions, as well as RNA recovered from bound fractions at rounds 2, 3 and 4, was amplified by emulsion PCR with two sets of primers as described previously [2]. Amplified DNA was purified by 2% agarose gel electrophoresis and sequenced using Illuminas HiSeq 2500 device with 100-cycle paired-end sequencing protocol.

6 Acknowledgments

We thank Rolf Backofen and Fabrizio Costa, Freiburg University for a stimulating discussion. This work was supported in part by the Intramural Research Program of the NIH, National Library of Medicine (JH, TMP), in part funds from the Laboratory of Hemostasis and the Center for Biologics Evaluation and Research, Food and Drug Administration's Modernization of Science program (ZES), and in part by bequest from the Dodson estate and the Sylvester Comprehensive Cancer Center, Medical School, University of Miami (AB and EG). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The findings and conclusions in this article have not been formally disseminated by the Food and Drug Administration and should not be construed to represent any Agency determination or policy.

References

1. Kim, Y.S., Gu, M.B.: Advances in aptamer screening and small molecule aptasensors. Advances in biochemical engineering/biotechnology (Jul 2013) PMID: 23851587.
2. Berezhnoy, A., Stewart, C.A., Mcnamara, James O, n., Thiel, W., Giangrande, P., Trinchieri, G., Gilboa, E.: Isolation and optimization of murine il-10 receptor blocking oligonucleotide aptamers using high-throughput sequencing. Molecular therapy: the journal of the American Society of Gene Therapy **20**(6) (Jun 2012) 12421250 PMID: 22434135.
3. Binning, J.M., Wang, T., Luthra, P., Shabman, R.S., Borek, D.M., Liu, G., Xu, W., Leung, D.W., Basler, C.F., Amarasinghe, G.K.: Development of rna aptamers targeting ebola virus vp35. Biochemistry (Sep 2013) PMID: 24067086.
4. Shi, H., Cui, W., He, X., Guo, Q., Wang, K., Ye, X., Tang, J.: Whole cell-selex aptamers for highly specific fluorescence molecular imaging of carcinomas in vivo. PloS one **8**(8) (2013) e70476 PMID: 23950940.
5. Cerchia, L., Hamm, J., Libri, D., Tavitian, B., de Franciscis, V.: Nucleic acid aptamers in cancer medicine. FEBS letters **528**(1-3) (Sep 2002) 1216 PMID: 12297271.
6. Macugen: Fda approves new drug treatment for age-related macular degeneration
7. Zichel, R., Chearwae, W., Pandey, G.S., Golding, B., Sauna, Z.E.: Aptamers as a sensitive tool to detect subtle modifications in therapeutic proteins. PloS one **7**(2) (2012) e31948 PMID: 22384109.
8. Ellington, A.D., Szostak, J.W.: In vitro selection of rna molecules that bind specific ligands. Nature **346**(6287) (Aug 1990) 818822 PMID: 1697402.
9. Hoinka, J., Zotenko, E., Friedman, A., Sauna, Z.E., Przytycka, T.M.: Identification of sequence-structure rna binding motifs for selex-derived aptamers. Bioinformatics (Oxford, England) **28**(12) (Jun 2012) i215223 PMID: 22689764.
10. Kupakuwana, G.V., Crill, James E, n., McPike, M.P., Borer, P.N.: Acyclic identification of aptamers for human alpha-thrombin using over-represented libraries and deep sequencing. PloS one **6**(5) (2011) e19395 PMID: 21625587.
11. Zhao, Y., Granas, D., Stormo, G.D.: Inferring binding energies from selected binding sites. PLoS computational biology **5**(12) (Dec 2009) e1000590 PMID: 19997485.

12. Ogawa, N., Biggin, M.D.: High-throughput selex determination of dna sequences bound by transcription factors in vitro. Methods in molecular biology (Clifton, N.J.) **786** (2012) 5163 PMID: 21938619.
13. Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J.M., Yan, J., Sillanp, M.J., et al.: Multiplexed massively parallel selex for characterization of human transcription factor binding specificities. Genome research **20**(6) (Jun 2010) 861873 PMID: 20378718.
14. Gionis, A., Indyk, P., Motwani, R. VLDB 99. In: Similarity Search in High Dimensions via Hashing. Morgan Kaufmann Publishers Inc. (1999) 518529
15. Couper, K.N., Blount, D.G., Riley, E.M.: Il-10: the master regulator of immunity to infection. Journal of immunology (Baltimore, Md.: 1950) **180**(9) (May 2008) 57715777 PMID: 18424693.
16. Andoni, A., Indyk, P.: Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. Commun. ACM **51**(1) (Jan 2008) 117122
17. Yang, K., Zhang, L.: Performance comparison between k-tuple distance and four model-based distances in phylogenetic tree reconstruction. Nucleic Acids Research **36**(5) (Mar 2008) e33 PMID: 18296485.
18. OpenMP Architecture Review Board: OpenMP application program interface version 3.0 (May 2008)
19. Schling, B.: The Boost C++ Libraries. XML Press (2011)

AptaCluster - A Method to Cluster HT-SELEX Aptamer Pools and Lessons from its Application Supplementary Materials

7 Supplementary Figures and Tables

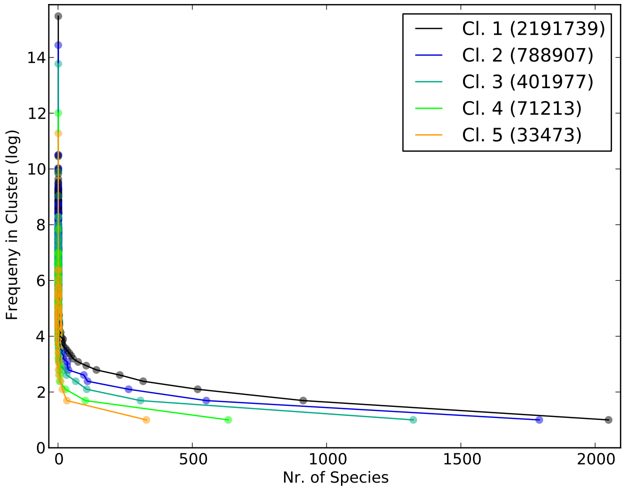

Fig. 4. The frequency distribution of the members of the 5 largest clusters. The cluster sizes are given in the brackets.

Table 1. Number of species with counts 1 to 5 present in the top 20 clusters of selection round 5 compared to the frequency of their occurrence in selection round 2. The overwhelming majority of the sequences are not present in the latter.

	Nr. of aptamers with frequency				
	1	2	3	4	5
Top 20, cycle 5	8529	2202	1074	614	465
Found in cycle 2	61	36	27	18	16